AI Course

Team Project Final Report

For students (instructor review required)

ⓒ2019 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of this document.

This document is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this document other than the curriculum of Samsung innovation Campus, you must receive written consent from copyright holder.

|  |
| --- |
| Fake News Detection |

**(14/8/2030)**

Intellibots

Hanan Alradadi

Haya Almossaeed

Rayan Dahlawi

Norah Alabbad

Content

1. Introduction

1.1. Background Information

1.2. Motivation and Objective

1.3. Members and Role Assignments

1.4. Schedule and Milestones

2. Project Execution

2.1. IoT Service Model

2.2. Data Processing

2.3. Service Implementation

2.4. System Diagram

3. Results

3.1. Data Acquisition (Sensor, Actuator, Controller)

3.2. Network & Communication

3.3. Hardware Implementation

3.4. Data Visualization

3.5. Testing and Improvements

4. Projected Impact

4.1. Accomplishments and Benefits

4.2. Future Improvements

5. Team Member Review and Comment

6. Instructor Review and Comment

1. Introduction

1.1. Background Information

The spread of false information has become a serious worry in today's information-driven culture. Media literacy and public awareness are directly impacted by the capacity to tell real news stories from fake ones. In order to create a fake news detection system, this project will make use of artificial intelligence and techniques for natural language processing.

1.2 Motivation and Objective

• The term "fake news" refers to misinformation or disinformation that has recently been spread through digital communication, such as WhatsApp messages, social media posts, etc. Fake news spreads more quickly than real news and causes issues and fear among various groups and in society.

We will classify whether a particular message or text is real or fake and handle these issues utilizing traditional NLP approaches.

• To pre-process the text and use several classification methods, utilize a Bag of n-grams.

• Implementations for Bag of Words are already incorporated into Sklearn CountVectorizer.

1.3 Members and Role Assignments

|  |  |
| --- | --- |
| Task | Member |
| Data cleaning | Rayan Dahlawi |
| Random Forest | Haya Almossaeed |
| Multinomial Naive Bayes | Norah Alabbad |
| KNN as Euclidean distance | Hanan Alradadi |
| Check the functionality of codes | All members |

1.4 Schedule and Milestones

|  |  |
| --- | --- |
| Task | Days |
| Gathering data | 2 |
| Prepare data | 3 |
| Analyze data | 3 |
| Train model and test model | 5 |
| Deployment | 5 |

2. Project Execution

2.1 Data Acquisition

The dataset used for this project is a collection of news items from Kaggle that have been classified as either fake or authentic.

Credits: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

2.2 Training Methodology

We use a machine learning pipeline technique to create an effective model. The procedure entails:

Data cleaning: Eliminating stop words, punctuation, and unhelpful words from the text data. In order to ensure uniformity in word forms, lemmatization is used.

Feature Extraction: Using the CountVectorizer method, convert the preprocessed text input into numerical representations. Trigrams, bigrams, and unigrams are thought to capture various context levels.

Model selection: Experimenting with various classifiers, such as Multinomial Naive Bayes, Random Forest, and K-Nearest Neighbors. The advantages and disadvantages of each classifier are assessed.

Model training: Using the preprocessed data to train the chosen classifiers while adjusting hyperparameters to maximize performance.

2.3 Workflow

The project workflow is divided into several phases:

Data Preprocessing: To prepare text data for feature extraction, it is cleaned, tokenized, and lemmatized.

The CountVectorizer extracts numerical features from text data while maintaining semantic meaning.

Model Training: To predict the veracity of news articles, a variety of classifiers are trained on the feature-transformed data.

Evaluation: Classification metrics like precision, recall, and F1-score are used to evaluate the performance of the model.

2.4 System Diagram

The following picture represents Architectural Design for Fake News Detection

|  |
| --- |
|  |

3. Results

3.1. Data Preprocessing

Data preparation is a crucial stage in ensuring the effectiveness of the system for identifying bogus news. It entails structuring unstructured text data so that it can be used for machine learning. The techniques used are as follows:

Text Cleaning and Transformation: Eliminating common terms, standardizing text, removing punctuation, and converting to lowercase maintain consistency.

Text Tokenization: Tokens (words) from the text are separated out for analysis and modeling.

While stemming trims words to their stem form, lemmatization reduces them to their base form.

Lemmatization and Stemming: While stemming trims words to their stem form, lemmatization reduces them to their base form.

Handling Rare and Common Words: To reduce noise and enhance discrimination, rare and excessively common words may be filtered out.

Vectorization: CountVectorizer: CountVectorizer converts text input into a numerical representation, producing a document-term matrix with word frequencies.

Handling Imbalanced Data: Oversampling or undersampling is used to correct for class imbalance and guarantee that both authentic and fraudulent news sources are fairly represented.

The quality of input data is optimized through effective data preparation, which enables the fake news detection system to recognize and understand relevant patterns with accuracy.

3.2 Exploratory Data Analysis (EDA)

To comprehend the dataset and direct following procedures, exploratory data analysis (EDA) is a vital stage. EDA entails:

Dataset Overview: Understanding class distribution and any imbalances in the dataset.

Text Length Analysis: Examining the lengths of articles to identify true and false news.

Word Frequency Analysis: Finding frequently used words in each class using word frequency analysis.

Visualization and Interpretation: Finding outliers, relationships, and anomalies using visualizations.

3.3 Modeling

KNN with different distance metrics, Random Forest, and Multinomial Naive Bayes are a few attempts at model construction. A thorough understanding of model performance is provided by evaluation metrics including precision, recall, and F1-score. The following table shows a summary of the performance measures for each model:

|  |  |  |  |
| --- | --- | --- | --- |
| model | Accuracy | F1-Score (Fake) | F1-Score (Real): |
| (KNN) with Euclidean Distance | 0.95 | 0.95 | 0.95 |
| Random Forest Classifier | 0.99 | 0.99 | 0.99 |
| Multinomial Naive Bayes | 0.99 | 0.99 | 0.99 |

Based on the information in the table above Both the Random Forest Classifier and Multinomial Naive Bayes perform remarkably well based on the evaluation measures. For both false and real news classes, they attain high accuracy and balanced F1-scores, showcasing their efficiency in identifying complicated links and managing large-scale data. The Random Forest Classifier is a good contender because of its ensemble method and capacity for handling non-linear interactions. The final decision between these models, however, might be influenced by elements like scalability, interpretability, and computing efficiency. To ensure generalization performance, model selection may necessitate additional testing, cross-validation, and testing on external datasets. The Random Forest Classifier is chosen as the ideal model for fake news detection based on the available data due to its remarkable accuracy and balanced F1-scores for both classes.

3.4 User Interface (Interface).

In order to give consumers a simple platform where they can enter news articles and get real-time classification, a user interface has been created. The following pictures are the interface forFake News Detection.

|  |
| --- |
|  |
|  |
|  |

3.5. Testing and Improvements.

The period of testing and improvement was extremely important in honing the fake news detection model. Major actions included:

Initial Model Evaluation: Using the test dataset, the chosen Random Forest Classifier underwent a thorough evaluation. Its performance was evaluated using metrics such as precision, recall, F1-score, and accuracy.

Analysis of Results: The confusion matrix and precision-recall trade-off were among the evaluation's results that were analyzed. This analysis showed potential improvement areas.

Addressing Limitations: To improve the model's performance, strategies like hyperparameter tuning, alternative feature engineering, and prospective ensemble techniques were investigated.

Cross-Validation: Cross-validation was used to confirm the robustness and generalizability of the model to new data.

Iterative Refinement: The phase was carried out incrementally, allowing for alterations based on the analysis. The goal of each iteration was to improve the model's functionality.

Final Model Evaluation: An assessment of the improved Random Forest Classifier's performance in comparison to the original model was conducted.

The model's dependability, accuracy, and capacity to handle unseen data were all guaranteed during the testing and refinement process, hence boosting its efficacy for fake news identification.

4. Projected Impact

4.1. Accomplishments and Benefits

For news consumers, educators, and researchers, the successful development of the fake news detection system is an invaluable resource. The project strives to create a society that is more informed by fostering media literacy and battling false information.

4.2 Future Improvements

To further improve accuracy and adaptability, future generations of the system might investigate more sophisticated NLP approaches, ensemble methodologies, and larger data sources. In addition deploy the modeling as web application.

5. Team Member Review and Comment

|  |
| --- |
| <ATTACH A TEAM PICTURE HERE> |

|  |  |
| --- | --- |
| NAME | REVIEW and COMMENT |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

6. Instructor Review and Comment

|  |  |  |
| --- | --- | --- |
| CATEGORY | SCORE | REVIEW and COMMENT |
| IDEA | \_\_/10 |  |
| APPLICATION | \_\_/30 |  |
| RESULT | \_\_/30 |  |
| PROJECT MANAGEMENT | \_\_/10 |  |
| PRESENTATION & REPORT | \_\_/20 |  |
| TOTAL | \_\_/100 |  |